Advanced Regression Ridge and Lasso Assignment Subjective Questions - Answers

Question 1

What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Ans.:

- Optimal value for alpha:
  - Ridge – 8
  - Lasso – 0.001

- After doubling both the alpha:
  - Ridge
    - Training R2 score decreases by a bit and test R2 score increases by a bit
    - Train RSS increases and test RSS decreases
    - Train and test MSE remain same
  - Lasso
    - Training R2 score decreases by a bit and test R2 score increases by a bit
    - Train RSS increases and test RSS decreases
    - Train and test MSE remain same

- Most important predictor variable for Ridge and Lasso before after doubling alpha is same i.e. OverallQual_9

Question 2

You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

Ans.:

We use Lasso regression in this case because for a model having such a high number of features, feature selection becomes important and Lasso does that by equating the coefficients of many features to zero

Question 3

After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

Ans.:

- Top 5 important variables in Lasso are as below:
  - OverallQual_9
  - OverallCond_9
  - OverallQual_8
  - GrLivArea
  - Neighborhood_Crawfor

- Top 5 variables after creating another model where the above features are not included are as below:
  - Condition2_PosA
  - 2ndFlrSF
  - Exterior1st_BrkFace
  - Functional_Typ
  - Neighborhood_Somers

<u>Question 4</u>

How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

Ans.:

As Per, Occam's Razor— Between two models that having similar performance in the limited training or test data, we should pick the one that makes fewer on the test data due to following reasons: -

▪ Simpler models are usually more 'generic' and are more widely applicable
▪ Simpler models require fewer training samples for effective training than the more complex ones and hence are easier to train.
▪ Simpler models are more robust.

o Complex models tend to change wildly with changes in the training data set
o Simple models have low variance, high bias and complex models have low bias, high variance
o Simpler models make more errors in the training set. Complex models lead to overfitting — they work very well for the training samples, fail miserably when applied to other test samples

Therefore, to make the model more robust and generalizable, make the model simple but not simpler which will not be of any use.

Regularization can be used to make the model simpler. Regularization helps to strike the delicate balance between keeping the model simple and not making it too naive to be of any use. For regression, regularization involves adding a regularization term to the cost that adds up the absolute values or the squares of the parameters of the model.